September 26, 2024

Iran’s AI-driven Social Media Botnets Target US Voters

Radware’s cyber threat intelligence (CTI) team has observed significant advancements in Iranian influence operations throughout 2024. It assesses with high confidence that the scale of these operations will significantly grow through the upcoming U.S. elections. These developments pose increased risks to the integrity of these elections, necessitating a reevaluation of current misinformation and disinformation defense strategies.

Key Attack Insights:

1. As part of its foreign policy, Iran attempts to manipulate public opinion through automated bots that spread misinformation by pretending to be real social media influencers.
2. As U.S. elections approach, these influence operations are expanding from targeting Israelis to focusing on Americans (see Figure 1).
3. Iranian actors are leveraging advanced AI technologies to generate convincing content, including deep fake videos and multilingual posts, significantly increasing the scale and sophistication of their campaigns.
4. The operations employ a cross-platform strategy, establishing presence across multiple social media services to increase credibility and reach, making containment and detection more challenging.
5. Iranian bot operators employ sophisticated TTPs, including the use of proxy IPs, AI-generated content and profiles, cross-platform coordination, and exploitation of current events, making their operations increasingly difficult to detect and counter.

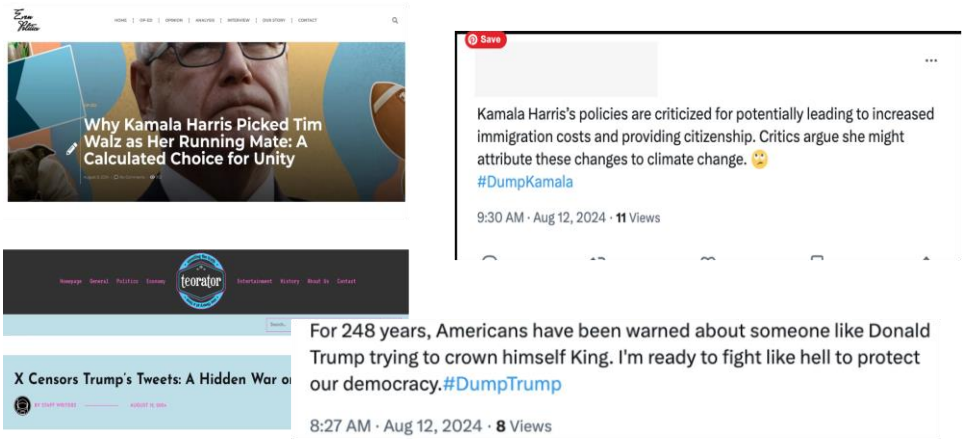


Figure 1: Examples of AI-generated fake news sites and social media botnet tweets captured by OpenAI.

Bot Operations TTP (Tactics, Techniques, and Procedures)

Operation architecture:

- Iranian operators employ specialized proxy services that make their bot traffic appear to originate from residential homes and mobile devices within the targeted country. This technique allows the bots to mimic the online behavior of genuine local users, significantly increasing the difficulty of detection and enhancing the credibility of their influence operations.

Account Management and Structure: Use of a tiered account system

- High-quality "avatar" accounts for direct interaction and content creation
- Larger numbers of lower-quality accounts or bots for amplification and engagement
- Fake accounts to manage Pages and post content [2] [4]

Content Generation and Manipulation: AI-powered content creation

- Use of AI models to generate and rewrite comments in multiple languages
- Creation of AI-generated newsreaders on YouTube
- Rapid adaptation of content to current events and emerging issues [1] [3]

Cross-Platform Tactics: Multi-platform presence and coordination

- Operation of fake news websites posing as both progressive and conservative outlets
- Coordinated posting and sharing across multiple social media platforms
- Driving traffic from social media to external websites [1] [4]

Engagement and Recruitment: Direct user interaction

- Messaging real users for recruitment or manipulation
- Collecting personal details through fake "volunteer" forms [2]

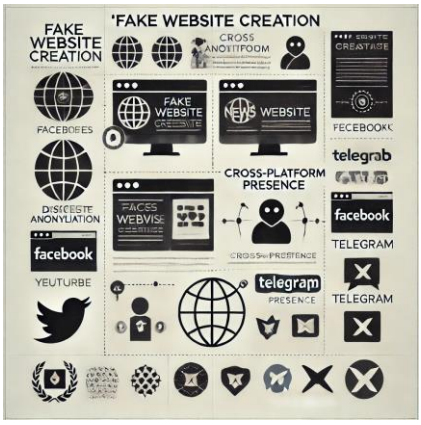
Amplification and Visual Manipulation:

- Use of bots to amplify specific viewpoints (e.g., pro-Russia views on Ukraine war) [3]
- Creation of AI-generated profile photos for fictitious personas using Generative Adversarial Networks (GANs) [3]

Public Opinion Manipulation Operation (Step by Step)

Part 1: Pre-attack Preparation

- **Account Acquisition:**
  - Create new social media accounts using bots or purchase social media accounts with established credibility.
  - Use AI-generated profile photos for fictitious personas, often created using Generative Adversarial Networks (GANs) [3].

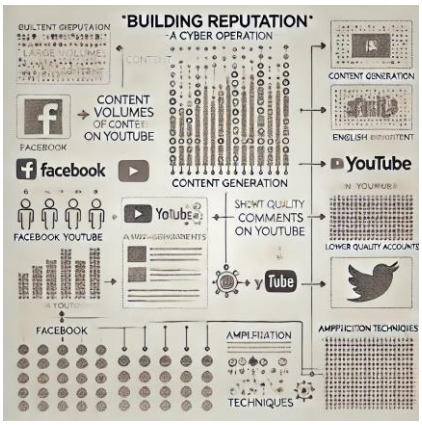


- **Fake Website Creation:**
  - Establish fictitious news websites with distinct branding [3].

- Use anonymization techniques to hide the true origin of these sites.

➢ **Cross-Platform Presence:**

- Establish a presence for fake entities across multiple internet services, including Facebook, YouTube, Telegram, and Twitter (now X) [3].



Part 2: Building Reputation

➢ **Content Generation:**

- Use AI tools to generate large volumes of non-political content [3].
- Create AI-generated newsreaders on YouTube [3].
- Generate short comments in English and Spanish using AI models [1].

➢ **Amplification Techniques:**

- Use a network of lower-quality accounts or bots for amplification [2].
- Engage in coordinated posting and sharing across multiple platforms.

➢ **Local Engagement:**

- Attempt to engage with local audiences through targeted content.

Part 3: The Attack

- **Disinformation Dissemination:**
  - Publish AI-generated articles on U.S. politics and global events on websites posing as progressive and conservative news outlets to gain and establish digital reputation and credibility [1].
  - Use AI-generated newsreaders on YouTube to criticize specific politicians or policies [3].
- **Exploitation of Current Events:**
  - Rapidly adapt narratives to exploit emerging issues or crises.
- **Targeted Messaging:**
  - Tailor content for specific audiences to maximize impact.
- **Deep Fake Deployment:**
  - Create and disseminate deep fake videos of political figures.
- **Cross-Platform Amplification:**
  - Share and amplify content across multiple platforms to increase reach and credibility.
  - Use the network of bots and fake accounts to engage with and further spread the disinformation.

“Tears of War” Case Study

The "Tears of War" network, as described in a recent report from Israeli think tank The Institute of National Security Studies, is an Iranian influence operation that emerged after the outbreak of the "Swords of Iron" war (which began on October 7, 2023). Here are the key details:

1. **Purpose:** The network's primary goal was to maintain and curate a database of Israeli casualties, including those killed, injured and taken hostage during the conflict.
2. **Content:** They regularly published visually designed content featuring images and details of Israeli victims.
3. **Platform presence:** The network maintained a coordinated presence across multiple social media platforms, including TikTok, Instagram, Twitter, Facebook and WhatsApp.
4. **Emotional manipulation:** By focusing on Israeli casualties and losses, the network aimed to demoralize the Israeli public and potentially influence public opinion regarding the ongoing conflict.
5. **Data collection:** The network also attempted to gather personal information from Israelis by encouraging them to participate in activities related to the hostages, such as hanging posters in public spaces.
6. **Fake job offers:** They posted "job offers" on Israeli job search platforms and groups, likely to collect more personal data or recruit unwitting participants.

Five Key Attack Trends

Our analysis reveals five key trends that characterize the latest generation of Iranian bot networks and influence operations:

1. AI-Powered Content Generation
Iranian actors increasingly leverage AI tools to generate large volumes of content for their influence operations. OpenAI identified and removed a cluster of ChatGPT accounts that generated content focused on multiple topics, including commentary on candidates on both sides of the U.S. presidential election [1]. This AI-generated content was then shared via social media accounts and websites, significantly increasing these operations' scale and potential impact.

The sophistication of AI-generated content is particularly concerning. For instance, Iranian groups’ use AI to create deep fake videos of political figures. In one case, a fake video of Israeli Prime Minister Netanyahu addressing global Jewry in English was created and disseminated

[2]. This level of sophistication in content creation could easily be applied to U.S. political figures, swaying voter opinions through seemingly authentic but entirely fabricated statements.

2. Cross-Platform Presence

To increase credibility and reach, Iranian actors establish a presence for their fake entities across multiple internet services. Microsoft reported that fictitious news brands had accounts on Facebook, YouTube, Telegram and Twitter (now X) [3]. This multi-platform approach helps these entities appear more legitimate and withstand scrutiny by platforms and researchers.

One striking example is the creation of a fake UK-based news outlet called "Euro Top News." This fictitious entity went to great lengths to appear legitimate, even listing the address of an actual publication, Ok! Magazine, and appropriating the VAT number from an unconnected UK-based digital marketing firm [3]. Such attention to detail in creating false legitimacy could easily be replicated for U.S.-focused disinformation campaigns.

3. Targeted Messaging for Specific Audiences

Iranian bot networks tailor their content for specific audiences to maximize impact. For English-speaking audiences, particularly in the U.S., the campaigns post primarily about diminishing support for Ukraine in the West. They create AI-generated newsreaders on YouTube that criticize U.S. President Biden and Democrats for providing aid to Ukraine instead of investing in their own country [3].

The ability to create targeted content extends to multiple languages. OpenAI found that the Iranian operation was creating short comments in both English and Spanish, which were posted on social media. Some of these comments were generated by asking AI models to rewrite comments posted by other social media users [1]. This multilingual approach allows Iranian actors to target diverse demographics within the US electorate.

4. Rapid Exploitation of Current Events

Iranian actors quickly adapt their bot networks to exploit current events and divisive issues. Microsoft reported that Iranian operations are likely to blame economic hardships in the U.S. on providing financial help to Ukraine, paint Ukraine's government as unreliable, or amplify voices expressing pro-Russia views on the war and its prospects [3].

This rapid adaptation was demonstrated in the past when Iran created bots, fake news sites, and social media profiles to exploit geopolitical tensions. Similar tactics could be employed to exploit any emerging controversies or crises in the lead-up to the U.S. elections.

5. Coordinated Inauthentic Behavior

Meta identified and removed 76 Facebook accounts, 30 Pages, and 11 Instagram accounts for engaging in coordinated inauthentic behavior originating from Iran. These networks used fake accounts to manage Pages, post content, and drive people to off-platform websites [4]. Some fake accounts were detected and disabled by automated systems before manual investigation,

highlighting the ongoing cat-and-mouse game between platform security and Iranian operators.

Israeli-targeted Tactics Now Used on Americans

Ascending timeline of Iran social media botnet operation to manipulate Israeli and American public opinion (Source: 1-4: INSS, 5-6: Meta, 7- OpenAI):

Reasons for Concern

Date	Target	Operation	Goal
October 2023	Israeli public	"Lion of Judah News" network spreads false information about "terrorists" in Israeli hospitals	Public opinion manipulation and social unrest
October 2023 - April 2024	Israelis, particularly those concerned with the hostage situation	"BringHomeNow" network impersonates hostage support groups, collects data, manipulates public discourse, and conducts real-world actions	Public opinion manipulation, data collection, and potential policy influence
October 2023 - July 2024	Israeli public	"Tears of War" network curates casualty database, publishes victim content, and attempts data collection across multiple social media platforms	Data collection demoralizes and influences public opinion regarding the ongoing conflict.
October 2023 - July 2024	Israelis, global Jewish community	"Israel Second" network operates across multiple social media platforms, spreads deep fake videos of Israeli Prime Minister Netanyahu and deep fake videos of five Israeli rabbis criticizing Netanyahu.	Public opinion manipulation, social division, and political destabilization
Early 2024	English-speaking audiences globally	Creation of a fictitious "Euro Top News" outlet	Disinformation dissemination and credibility building
Early-Mid 2024	Americans	AI-generated newsreaders on YouTube criticize Biden and Democrats	Political influence and public opinion manipulation
June-August 2024	Americans, Spanish-speaking Americans	Removal of a cluster of ChatGPT accounts used to generate and disseminate content for the U.S. presidential election, including short comments in English and Spanish and articles on U.S. politics published on fake news websites	Election interference, political influence, and public opinion manipulation

- AI-powered content generation could lead to a surge in high-quality, persuasive disinformation that is difficult to distinguish from legitimate content. The ability to create

deep fake videos of political figures poses a particular threat to informed decision-making by voters.

- Cross-platform presence makes it challenging to contain the spread of disinformation and increases the potential for viral distribution. Creating seemingly legitimate news outlets with a presence across multiple platforms provides a veneer of credibility that can be difficult for average users to see through.

- Targeted messaging tailored to specific audiences may exacerbate societal divisions and polarize the electorate. Iranian actors' ability to generate content in multiple languages allows them to target diverse demographics within the US population, potentially amplifying divisions along ethnic or linguistic lines.

- Rapid adaptation to current events allows for timely exploitation of emerging issues, potentially swaying public opinion in critical moments. The ability to quickly pivot narratives in response to global events suggests that Iranian actors could rapidly exploit any controversies or crises that emerge during the election campaign.

- Coordinated inauthentic behavior complicates detection and mitigation efforts, potentially overwhelming existing defense mechanisms. The use of AI-generated profile photos and sophisticated persona-creation techniques makes it increasingly difficult to distinguish between genuine and fake accounts.

- The technical sophistication of these networks is evident in their use of AI-generated profile photos for their fictitious journalist personas. These consistent profile photos across different internet platforms are often created using Generative Adversarial Networks (GANs), making the fake personas appear more convincing [3]. This level of detail in creating fake personas significantly increases the challenge of detecting and countering these influence operations.

Resources List:

[1] OpenAI, "Disrupting a covert Iranian influence operation," August 16, 2024, https://openai.com/index/disrupting-a-covert-iranian-influence-operation/

[2] INSS, "Iranian Foreign Intervention and Influence During the Swords of Iron War", August 11, 2024]

https://www.inss.org.il/social_media/the-report-that-exposes-iran-has-significantly-increased-its-activities-in-the-cyber-realm-against-israel/

https://www.inss.org.il/he/publication/iranian-influence/

[3] Microsoft, "Iran steps into US election 2024 with cyber-enabled influence operations", August 9, 2024, https://www.microsoft.com/en-ie/security/security-insider/intelligence-reports/iran-steps-into-us-election-2024-with-cyber-enabled-influence-operations

[4] Meta, "Second Quarter Adversarial Threat Report", August 2024, https://www.politico.eu/wp-content/uploads/2023/08/29/NEAR-FINAL-DRAFT-Meta-Quarterly-Adversarial-Threat-Report-Q2-2023.pdf

EFFECTIVE DDOS PROTECTION ESSENTIALS

Behavioral-Based Detection – Leverage Radware's advanced behavioral analysis to quickly and accurately identify and block anomalous bot activity while allowing legitimate traffic.

Real-Time Signature Creation – Utilize Radware's ability to promptly create and deploy signatures to protect against emerging threats and zero-day attacks.

AI-Powered Content Analysis – Implement Radware's AI-driven solutions to detect and mitigate sophisticated disinformation campaigns across multiple platforms.

Cross-Platform Monitoring – Employ Radware's comprehensive monitoring tools to track influence operations across various digital channels.

Rapid Response Capabilities – Leverage Radware's 24/7 Emergency Response Team to swiftly address and mitigate emerging threats.

For further network and application protection measures, Radware urges companies to inspect and patch their systems to defend against risks and threats.

EFFECTIVE WEB APPLICATION SECURITY ESSENTIALS

Full OWASP Top 10 coverage against defacements, injections, etc.

Low false positive rate using negative and positive security models for maximum accuracy

Auto-policy generation capabilities for the widest coverage with the lowest operational effort

Bot protection and device fingerprinting capabilities to overcome dynamic IP attacks and achieve improved bot detection and blocking

Securing APIs by filtering paths, understanding XML and JSON schemas for enforcement, and using activity tracking mechanisms to trace bots and guard internal resources

Flexible deployment options including on-premises, out-of-path, virtual or cloud-based

LEARN MORE AT RADWARE'S SECURITY RESEARCH CENTER

To know more about today's attack vector landscape, understand the business impact of cyberattacks, or learn more about emerging attack types and tools, visit Radware's **Security Research Center**. Additionally, visit Radware's **Quarterly DDoS & Application Threat Analysis Center** for quarter-over-quarter analysis of DDoS and application attack activity based on data from Radware's cloud security services and threat intelligence.

PARTICULAR PURPOSE OR NON-INFRINGEMENT. USE OF THIS REPORT, IN WHOLE OR IN PART, IS AT USER'S SOLE RISK. RADWARE AND/OR ANYONE ON ITS BEHALF SPECIFICALLY DISCLAIMS ANY LIABILITY IN RELATION TO THIS REPORT, INCLUDING WITHOUT LIMITATION, FOR ANY DIRECT, SPECIAL, INDIREC, INCIDENTAL, CONSEQUENTIAL, OR EXAMPLARY DAMAGES, LOSSES AND EXPENSES ARISING FROM OR IN ANY WAY RELATED TO THIS REPORT, HOWEVER CAUSED, AND WHETHER BASED ON CONTRACT, TORT (INCLUDING NEGLIGENCE) OR OTHER THEORY OF LIABILITY, EVEN IF IT WAS ADVISED OF THE POSSIBILITY OF SUCH DAMAGES, LOSSES OR EXPENSES. **CHARTS USED OR REPRODUCED SHOULD BE CREDITED TO RADWARE**